Search Strategies for IR Spectra - Part II Recognizing Problems and Improving the Intelligence of the Search Algorithms

Author: Michael Boruta

This Application Note will address common questions Bio-Rad receives in regards to IR database searching. Bio-Rad's purpose in doing so is to try to help you with your selection of algorithms and interpretation of your results. This Application Note will explain techniques used to increase the intelligence of the search algorithms.

Before trying to improve the intelligence of a search algorithm, one must first realize why the algorithm failed to give us the desired results. Although the existing full spectrum search algorithms often perform well without any modifications, there are cases where the results can be confusing or disappointing.

In the case where the unknown spectrum does not exist in the reference database or the unknown spectrum is a mixture, the presence of a band in the unknown containing a large percentage of the total area of the spectrum can skew the results of a search. For example, a strong and broad to medium width band, such as an OH or CH stretch, could cause the search algorithm to weigh the results very heavily toward compounds containing that band. This will minimize the contributions of the other more discriminating bands in the spectrum. If the unknown spectrum is not in the reference database, this could make it difficult to classify the unknown spectrum. If the unknown spectrum is a mixture, this could make it difficult to identify the unknown spectrum.

Recognizing the case where the search results do not match well with the unknown and the matches are skewed toward a large area band is the first step to improving the "intelligence" of the search algorithm. Often, limiting the range of the full spectrum search to exclude the large area band can help to better classify or identify an unknown spectrum.

In the following example, the "unknown" spectrum is a test mixture of cyclohexane and benzene. This spectrum exhibits a large area band in the CH region. Figure 1 shows the results of a full spectrum search using the unknown against the 200 compound Sadtler SR demo database. The first hit is shown above the unknown spectrum. Notice, not only are the Hit Quality Indices (HQI) low numbers, but the break or difference between successive HQIs is small. The Hit Quality Index is a measure of how well an unknown spectrum matches a reference spectrum.

This lack of a break is a strong indication that the search algorithm was not able to determine a good match. If we were to look quickly at the first several matches, the predominant matching feature would be the CH stretching region around 2900 cm^{-1}.

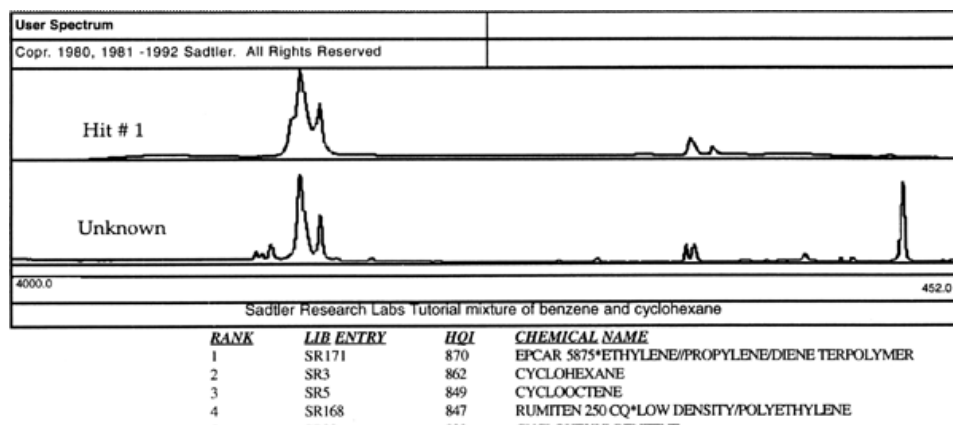| RANK	LIB ENTRY	HQI	CHEMICAL NAME
1	SR171	870	EPCAR 5875*ETHYLENE//PROPYLENE/DIENE TERPOLYMER
2	SR3	862	CYCLOHEXANE
3	SR5	849	CYCLOOCTENE
4	SR168	847	RUMITEN 250 CQ*LOW DENSITY/POLYETHYLENE
5	SR96	833	CYCLOHEXYL BENZENE

Figure 1: Full spectrum euclidean search of the test spectrum against the 200 compound SR demo database.

Figure 2 shows the results of a search in which the search range was limited to 3700-3000 and 2800-500 cm^{-1}. This excludes the large area band which was dominating the search results in the full spectrum search. Again, none of the HQI's are above 900, but the break in the HQI between the 1st and 2nd hits is now 245. This large break is a very strong indicator of a good match.

This example shows a case where excluding a large area band forced the search algorithm to focus on the rest of the spectrum. We were then able to confidently identify a component of a mixture even though it seemed as if the search algorithm had failed when the full spectrum was searched. This technique of finding a large area band and eliminating it from consideration during the search can have positive benefits when trying to classify unknowns or identify mixtures.

No single technique seems to work in all cases. However, it is generally best to start with a full spectrum search in order to emphasize the large area bands in the unknown spectra. If this fails to produce the desired results, you can attempt to modify the search parameters to improve the results.
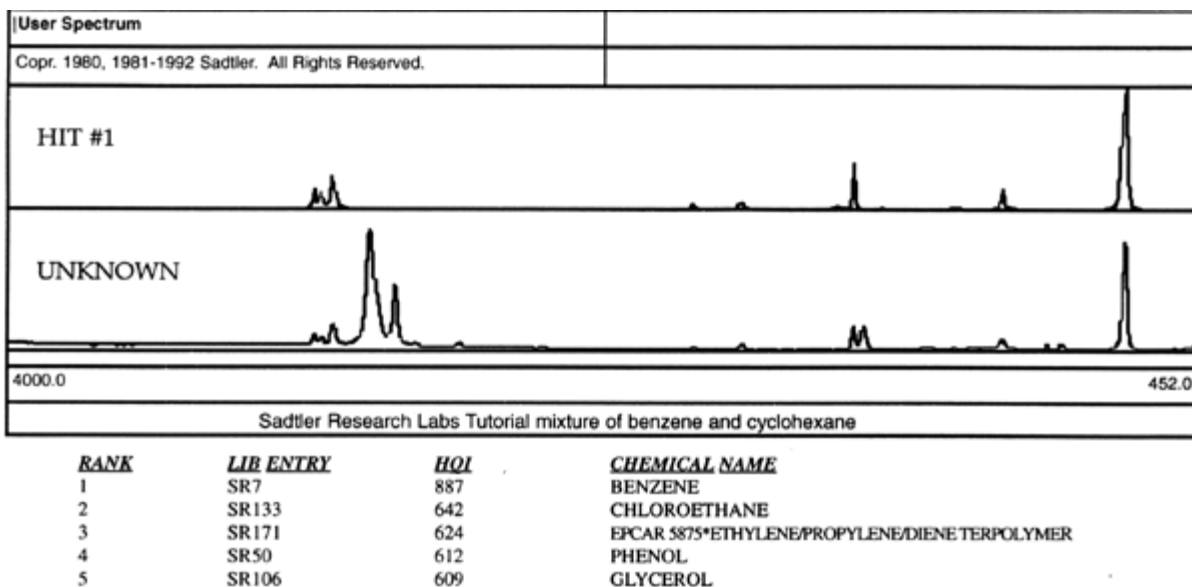

RANK	LIB ENTRY	HQI	CHEMICAL NAME
1	SR7	887	BENZENE
2	SR133	642	CHLOROETHANE
3	SR171	624	EPCAR 5875*ETHYLENE/PROPYLENE/DIENE TERPOLYMER
4	SR50	612	PHENOL
5	SR106	609	GLYCEROL

Figure 2: Limited range (3700-3000 & 2800-500) spectrum euclidean search of the test spectrum against the 200 compound SR demo database.

BIO-RAD

Bio-Rad Laboratories

94035